

科技文献检索系统语义丰富化框架的设计与实践*

谢 靖 王敬东 吴振新 张智雄 王 颖 叶志飞
(中国科学院文献情报中心 北京 100190)

摘要:【目的】通过采用语义识别、知识关系计算等方法提升科技文献检索系统的服务功能和效果,使之能够呈现更加丰富的知识化语义信息,将更多的知识点和知识关系展现给用户。【方法】应用数据挖掘和关系计算工具,深度识别和抽取科技文献中的语义知识,分析、计算、构建语义关系,并将得到的语义知识和语义关系建立多维语义索引树,设计新的数据组织呈现模型。【结果】研发语义丰富化检索示范系统,在科技文献检索应用过程中充分揭示语义信息,丰富检索体验。【局限】选取的试验数据集合不够充足,缺少其他领域应用对比。【结论】本文模型设计给用户带来更多的知识层面的关联、揭示和导航,提升了检索系统体验。同时分析了设计模型的不足之处,探索改进方法。

关键词: 语义丰富化 语义知识组织 语义关系呈现 多维索引
分类号: TP391

1 引 言

随着语义技术、知识图谱和本体技术的迅速发展和在科技文献中的应用,如何发掘揭示科技文献之间的语义关系,充分利用知识和体现知识价值,是当今科技文献检索关注的重点。人们不再满足原有"关键词+检索列表"模式的检索系统,而希望通过语义途径检索发现科技文献,在科技文献的检索系统中呈现知识点和知识关系等更加丰富的内容。语义丰富化框架的设计目标是改进现有单一关键词导向的检索系统,将多种类型的语义知识[1],知识之间丰富的关联关系等深层信息,利用数据挖掘与呈现技术重新组织,在科技文献检索过程中充分揭示出来。

2 语义丰富化现状分析及研究意义

当前基于知识图谱[2]的语义搜索引擎,如 Google Knowledge Graph[3],利用知识图谱改进传统搜索引擎的呈现方式,分析用户输入生成关联的百科知识,辅助组织多类型语义知识及多媒体展现,很大程度提升了用户检索体验。知名的 WolframAlpha[4]和 Kngine[5]智能语义搜索引擎,更是将语义搜索展现为一种智能知识问答方式。在强大的百科知识库和知识图谱支持基础上,对用户输入问题智能解析、搜索并给出相关的答案。

知识图谱的搜索引擎仅对用户输入进行语义丰富化,揭示知识图谱中的既有知识,不能发现科技文献本身潜在的知识。在文献发现过程中依然采用传统检索架构,使用列表方式呈现相关文献。而 SindiceTech[6]平台的研究应用,实现了对文本数据的深度拆解、语义关系计算等智能方法,将海量文本数据全部用 RDF 三元组[7]方式表示,以发现文本中潜在知识为向导,形成

通讯作者:吴振新,ORCID:0000-0003-4966-1961,E-mail:wuzx@mail.las.ac.cn。

*本文系中国科学院文献情报能力建设专项"基于大数据计算的资源发现平台建设"(项目编号:院1676)和国家社会科学基金青年项目"基于关联数据的学术资源深度挖掘方法研究"(项目编号:15CTQ006)的研究成果之一。

FreeBase[8]关联知识库，开创了三元组搜索展示数据的先例。这种深度知识关系揭示方法对发现原始文本内部的潜在知识关系具有重要意义。

由于 SindiceTech 平台面向互联网广泛领域的数据采集、组织，没有针对特定科研领域的实体名称和关系进行规范和控制，检索产生较多噪声数据，影响了语义关系检索的效果，因此没有在科技文献检索中应用。本文设计思路综合了知识图谱检索和三元组数据组织发现两种方式，充分利用专业领域知识，并在科技文献深度标注和知识关系计算的基础上，对文献中出现的知识和关系进行规范。综合以上平台的设计思路，设计语义丰富化呈现模型。利用 Apache Solr[9]分面机制设计多维索引，充分发掘揭示既有语义关系和潜在语义关联，从而在用户输入端和检索过程中提升语义丰富化检索体验。

为体现科研领域主题范围内的语义丰富化效果，本文的试验选择 PubMed[10]的医学领域的 Migraine Disorder、Heart Diseases 这两个主题近两年内的文章集合作为示范系统的试验数据集，采用医学领域数据挖掘计算较为成熟的 SemRep[11]和 ClausIE[12]作为基础数据挖掘分析工具，研发了检索示范系统以探索科技文献检索的语义丰富化的效果。

3　语义丰富化的总体架构设计

如图 1 所示，科技文献检索系统语义丰富化的总体架构设计分为语义计算和语义索引两个部分。语义计算面向知识的挖掘与组织，语义索引面向知识的揭示与应用。

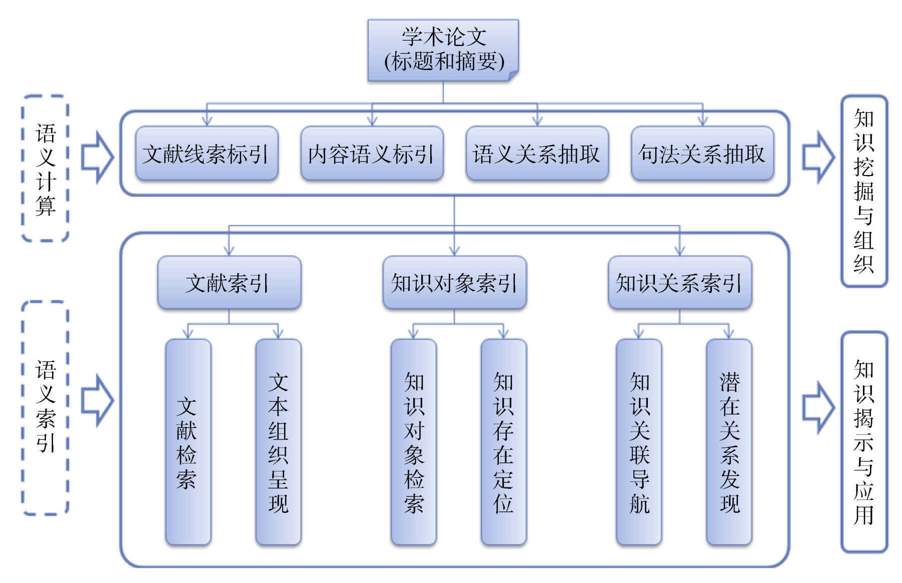


图 1　语义丰富化总体设计框架

3.1　语义计算

文本中包含的具有语义价值的术语和实体，本文统称为知识对象[13]。语义计算工作首先标识出文献的关键词，并识别出它所属的类型(即它是什么)，即得到本文需要的知识对象；其次计算出各个知识对象之间的关联关系。如图 1 上部分所示，语义计算包括：文献线索标引、内容语义标引、语义关系抽取、句法关系抽取。

(1) 文献线索标引：主要包括对文献基础描述元数据的加工、标引。根据文献提供的结构化摘要文本，切分研究目的、研究方法、研究工具、研究结果等结构化摘要，并实现句子切分预处理，以支持后续知识对象关系抽取。

(2) 内容语义标引：实现文本内容中具有语义价值的术语和实体的标引，通过内容标引过程得到试验需要的知识对象。

(3) 语义关系抽取：根据同一个句子中计算标引得到的知识对象，在 UMLS[14]和 STKOS 科技知识组织体系开放引擎[15]中查询每两个知识对象可能存在的语义关系[16]，并将语义关系记录为 S-P-O 三元组[17]。

(4) 句法关系抽取：使用自然语言处理方法计算句法关系[18]，将长句拆分成短句和子句，并在短句中计算出主谓宾关系，将主谓宾关系以 S-P-O 三元组的

方式记录。一个长句子可以拆分、记录为多个 S-P-O 三元组。

3.2 语义索引

语义索引工作将语义计算后得到的文献线索、文本内容、知识对象、对象关系等不同维度的数据，构建多维度的语义索引体系，将各个维度数据有机组织起来，便于语义丰富化检索系统揭示应用。如图 1 下半部分所示，语义索引分为三个部分:

(1) 文献索引: 对文章的基本描述元数据进行索引、文本切分后的句子片段索引，保障文章与所属句子片段的映射关系，用于文献元数据查询浏览和文本片段组织呈现。

(2) 知识对象索引: 用于对知识对象的检索查询和分类展示，识别并规范用户输入关键词。并将知识对象与文献文本组织关联，用于定位知识对象存在于某篇文献和具体某个句子之中。

(3) 知识关系索引: 语义标引工作得到语义、句法两种关系，文本索引将这两种关系合并统称为知识关系，它们均以 S-P-O 三元组的方式表达。知识关系索引以三元组为基础构建，实现知识关联导航和潜在知识关系发现。

4 语义计算框架

语义计算框架工作流程如图 2 所示，参照 MeSH 主题词表[19]将医学领域知识对象划分为 16 个一级大类、134 个二级分类[20]和 30 种谓词语义关系(根据 NLM 每年更新，分类和关系的数目、结构有所变动)，选取医学领域文献的标题和摘要文本数据作为试验素材，使用 SemRep 和 MetaMap[21]工具对试验文本数据中的重要知识对象进行标引抽取。使用 MetaMap 和 ClausIE 工具实现对试验文本数据中的语义关系计算识别。

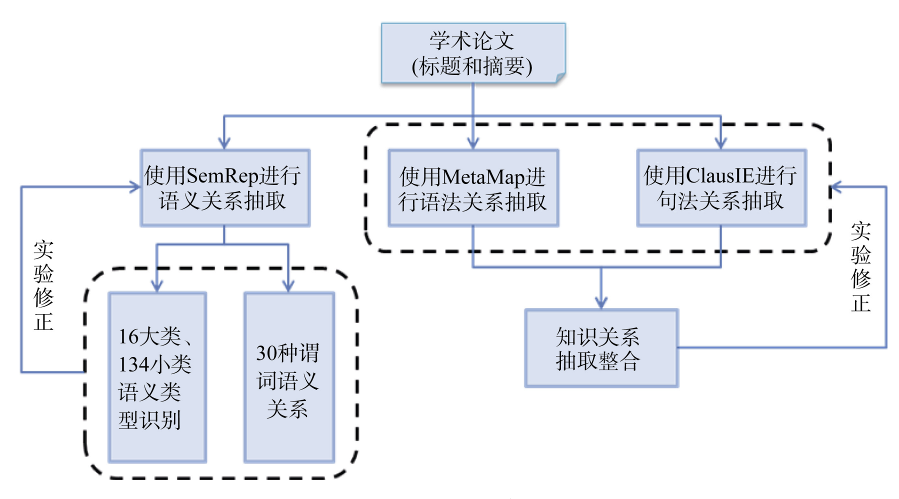


图 2 语义标引流程

4.1 内容语义标引

(1) 文献内容的语义标引

通过 StanfordTagger[22]和 ClausIE 句法解析工具对 20 万条科技文献进行分句等预处理; 利用 dTagger[23]对预处理后的科技文献进行词性标注; 通过美国国立医学图书馆基于 UMLS 超级叙词表开发的 MetaMap 工具将句子拆分成具有意义的短语片段。如对标题 "Effectiveness of behavioural management on migraine in adult patients visiting family practice clinics: a randomized controlled trial"处理，可得到 A-E，5 个短语，具体标引示例如图 3 所示。

标题:"Effectiveness of behavioural management on migraine in adult patients visiting family practice clinics: a randomized controlled trial."

MetaMap分析结果:

A:'Effectiveness of behavioural management' ; **B:**'on migraine in adult patients' ; **C:**visiting
D:'family practice clinics'; **E:**'a randomized controlled trial'

A:(lexmatch(effectiveness),inputmatch(Effectiveness), tag(noun));
(lexmatch(behavioural),inputmatch(behavioural),tag(adj));
(lexmatch(management),inputmatch(management),tag(noun))
........|
E:(lexmatch([a]),inputmatch([a]),tag(det));
(lexmatch('randomized controlled trial'),inputmatch(randomized,controlled,trial),tag(noun))

图 3 MetaMap 语义标引结果示例

(2) 知识对象的语义识别

语义标引得到的结果，通过 SemRep 工具与

UMLS 超级叙词表进行映射，并识别出各词的语义类型，以及抽取可信赖的语义关系。同图 3 示例，经 SemRep 对上述标题分析后可得到 10 个实体和语义关系，具体如表 1 所示。通过语义识别实现了文本标引内容与 MeSH 主题词表 16 个一级大类、134 个二级分类的重要知识对象的识别映射。

表 1　SemRep 知识对象语义识别结果

SemRep 标记	文章 PMID	来源标记	文本位置	术语类型	MeSH 词表术语代码	MeSH 词表标准术语	语义关系缩写	文本中原始词汇	置信度	术语开始位置	术语结束位置
SE	00000000	tx	1	entity	C1280519	Effectiveness	qlco	Effectiveness	1000	1	13
SE	00000000	tx	1	entity	C0150143	Behavior management	topp	behavioural management	964	18	39
SE	00000000	tx	1	entity	C0149931	Migraine Disorders	dsyn	migraine	1000	44	51
SE	00000000	tx	1	entity	C0001675	Adult	aggp	adult	888	56	60
SE	00000000	tx	1	entity	C0030705	Patients	podg	patients	888	62	69
SE	00000000	tx	1	entity	C0015607	family medicine (field)	bmod	family practice	901	81	95
SE	00000000	tx	1	entity	C0442592	Clinic	hcro,mnob	clinics	901	97	103
SE	00000000	tx	1	entity	C1514720	Randomized	ftcn	randomized	851	108	117
SE	00000000	tx	1	entity	C0702113	Controlled	ftcn	controlled	851	119	128
SE	00000000	tx	1	entity	C0008976	Clinical Trials	resa	trial	851	130	134

语义关系识别结果:
SE|00000000||tx|1|relation|3|1|C0149931|Migraine Disorders|dsyn|dsyn|||migraine|||1000|44|51|
PREP|PROCESS_OF||53|54|3|1|C0030705|Patients|podg,humn|humn||patients||||888|62|69

4.2　语义关系计算识别

利用 MetaMap 工具实现语义关系的计算识别，将知识对象关系识别为 30 个规范关系；利用 ClausIE 工具实现对句法树关系的识别；将 MetaMap 和 ClauseIE 两种工具识别的语义关系数据合并整合，参照 MetaMap 选取的 30 个规范关系对试验数据规范修正。完成的数据组织、规范工作包括：

(1) 实现文献内部知识对象的语义关系标引，通过 SemRep 和 MetaMap 工具实现科技文献中 30 种语义关系的抽取，挖掘知识对象之间潜在的知识关系。

(2) 实现文献内容的句法关系标引，通过 ClausIE 工具实现科技文献中句法关系(S-P-O)抽取，发现知识对象(关键词、术语)之间潜在关联关系。

(3) 整合语义关系和句法关系标引，对 1 116 篇文献摘要进行抽取，共提取 S-P-O 关系 50 204 条，包括语义关系 41 590 条，以及语法关系 8 614 条。

4.3　关键问题解决方案

(1) 标引内容与 MeSH 词表映射

SemRep 处理后的结果如表 1 所示，以第一行数据为例，红色字段(例子中的 qlco)为 134 个二级分类的语义关系缩写，本文完成了 MeSH 词表对应的 16 个一级大类、134 个二级分类的英文全称、英文缩写及中文名称收集整理。通过红色字段进行关联，建立起文本识别术语与 MeSH 词表映射关系。从而解决了 SemRep 处理后的结果与 MeSH 主题词表对应关系问题。

(2) ClausIE 抽取出的主语(S)、谓词(P)与 UMLS 超级词表的对应

ClausIE 按照句法关系抽取三元组与 SemRep 抽取的实体不能完全匹配；同时 SemRep 只能抽取语义动词，其他动词都被忽略掉。对于前一种情况通过模糊匹配的方式保证实体的对应；对于第二种情况，使用从 MetaMap 中提取出动词，然后进行匹配的方法，保障试验数据的规范性和一致性。

5　语义索引体系构建

语义索引设计目标是揭示知识对象和对象间多种语义关系，改变当前单一维度索引的方式，使用多棵索引树整合协同工作，多维度呈现语义内容。如图 4 所示，语义索引以知识对象为核心，遵循用户使用流程，从检索关键词出发，通过知识对象索引对输入关

键词进行语义识别和语义消歧；然后通过知识对象关系索引，遍历知识网络，导航、筛选所需关联知识；通过桥接索引确定知识对象所在的句子、段落；最后通过文献索引查询、展示包含相关知识内容的文献信息。

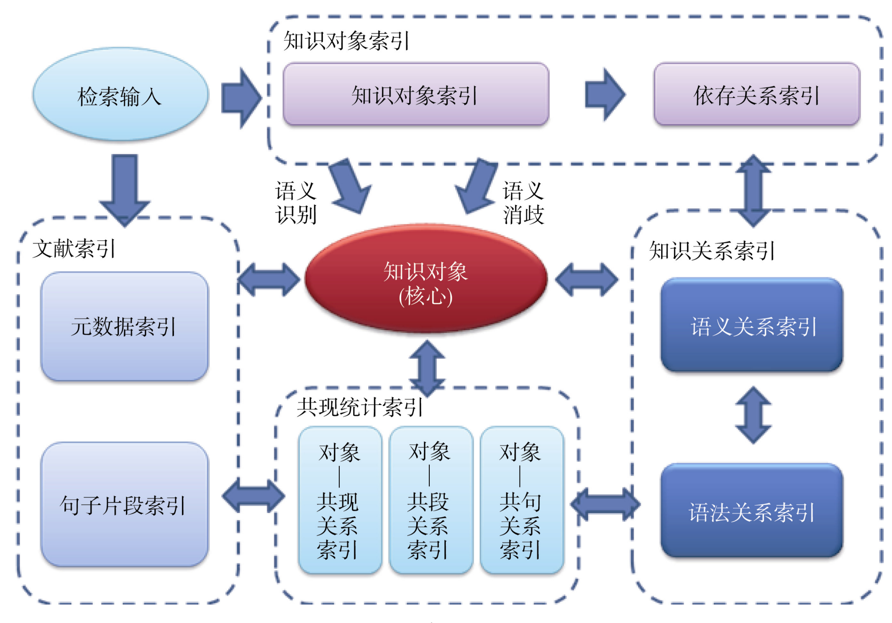


图 4　语义索引架构设计图

5.1　语义索引的功能与组织结构

基于上述4个步骤，将索引分为4个功能部分：

(1) 知识对象索引

①知识对象索引：将用户检索输入的关键词转换为相关知识对象，实现语义检索转变。检索并展示知识对象的各项属性，发现语义冲突的关键词，实现语义消歧功能。

②依存关系索引：索引知识对象存在的位置(如文章具体句子)，用于快速查询定位包含该知识点的文献。

(2) 知识关系索引

①语义关系索引：索引文本中出现的知识对象间的语义关系(上述语义计算得到的语义关系，利用STKOS知识组织系统规范语义关系[24])，实现语义关系的检索和分析展示功能。

②语法关系索引：索引文本中出现的知识对象间的语法关系(语法关系是基于NLP[25]句法分析得到句法关联关系)，用于区别语义和语法关系的检索和分析展示。

(3) 共现统计索引

①对象–共现关系索引：实现知识对象和存在同一文献的索引记录，用于分析和揭示语义知识对象的共现关系。

②对象–共段关系索引：实现知识对象和存在同一段落的索引记录，用于分析和揭示语义知识对象的共段落关系。

③对象–共句关系索引：实现知识对象和存在同一句子的索引记录，用于分析和揭示语义知识对象的共句子关系。

(4) 文献索引

①元数据索引：索引文献的元数据描述信息，用于文献描述信息的展示。

②片段句子索引：对文章的文本内容和句子索引，用于展示文献内容和相关知识对象的高亮显示等功能。

本文试验根据选取的数据集，共实现索引文献1 116篇，段落4 023个，句子7 684个，索引知识对象4 935条。索引知识关系50 204条。

5.2　关键问题及解决方案

(1) 输入关键词与知识对象的映射

试验中可能出现用户输入关键词语与索引中知识对象不能完全匹配问题，造成无法映射到准确的知识对象的问题；输入的一个关键词可能包含多种含义，发生语义识别歧义，无法明确映射到具体知识对象。

对第一个问题，采用索引模糊匹配方法，选取匹配分值最高的知识对象，并列出匹配的前5条列表通知用户，以便再次人工修正语义识别；对第二个问题，则提示用户存在不同含义的知识对象，由用户选择修正实现语义消歧。

(2) 知识对象之间关联关系的统计揭示

知识对象之间的关系都以三元组S-P-O的方式在Apache Solr中建立索引，如何使用Solr检索分面机制统计揭示知识对象之间的关联关系是一个难题。本文采用在三元组索引中加入冗余字段的方法，索引结构如表2所示。当检索主语(S)时对谓语宾语(P+O)组合字段分面，检索宾语(O)时对主语谓语(S+P)组合字段

表 2　三元组索引字段描述表

索引字段	字段描述	字段功能
S	三元组主语	检索查询
P	三元组谓语	检索查询
O	三元组宾语	检索查询
S+P	主语与谓词拼接组合	分面揭示
P+O	谓词与宾语拼接组合	分面揭示

分面。利用 Solr 的分面和频次统计功能，在检索任意知识对象时，通过对(P+O)和(S+P)分面，即可在检索结果集中揭示出现频次 TopN 的关联知识对象，帮助发现潜在知识。

6　语义丰富化示范平台实现

6.1　试验系统的数据组织

为实现语义丰富化检索示范平台，系统将数据组织为 4 个维度，如图 5 所示。

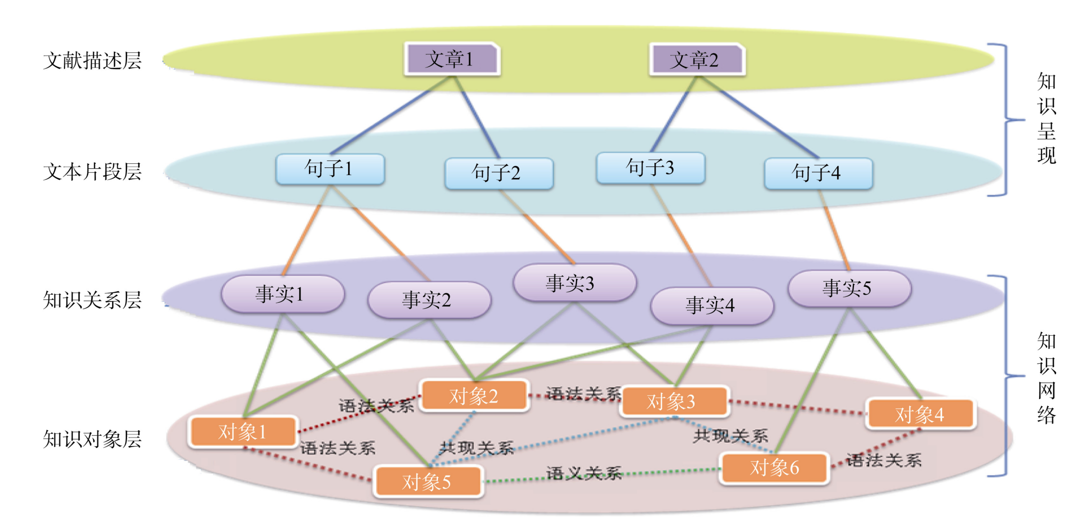


图 5　语义丰富化检索的数据组织结构

第一维度是文献描述层，对文章的标题、作者、摘要、发表时间等基本元数据表达揭示；第二维度是文本片段层，将文本摘要切分成段落和句子，对句子和段落关联揭示；第三维度是知识关系层(图 5 中将知识对象及它们之间的关系统称为事实)，用于对知识关联关系表达揭示；第四维度是知识对象层，对文本中识别出来的知识对象表达揭示。

从下而上的视角看，第三、第四维度将科技文献拆分为知识对象和知识关联关系，形成语义丰富化的科技文献知识网络，用于语义化的查询与关联导航。第一、第二维度结合文献的基本信息和文本片段，将知识与文献有机关联组织，用于知识呈现，辅助文献检索阅读。

6.2　示范平台功能实现

语义丰富化示范平台围绕用户的知识化应用需求，用户检索流程设计为：输入识别诠释、知识关系展示、潜在知识关联发现、语义辅助浏览。语义丰富化示范平台的研发实现了这 4 个功能。

(1) 识别诠释用户输入

示范系统第一步根据用户输入关键词识别出相应的知识对象，诠释并呈现具体知识内容。如图 6 所示，输入检索关键词“Headache”，系统识别“Headache”相关的知识对象，它是属于“体征或症状”的类型范畴。同时给出关于 Headache 的百科词条解释和相关的图片，并列出了“Headache”相关的知识对象以便用户选择修正。



图 6　语义识别功能展示

较之传统文献检索，该功能可以规范用户输入，将简单的关键词匹配检索转变为具有语义特征的知识对象检索，使得语义丰富化检索更加精准。同时语义识别功能可以标示知识对象的所属类型，辅助用户进行语义消歧。避免传统关键词检索经常出现的语义偏差。

(2) 知识关系图形展示

知识关系揭示功能以检索输入的知识对象为中心，在检索结果中以图的方式揭示了与其相关的知识对象、知识关系、知识所在的文章片段，如图 7 所示。

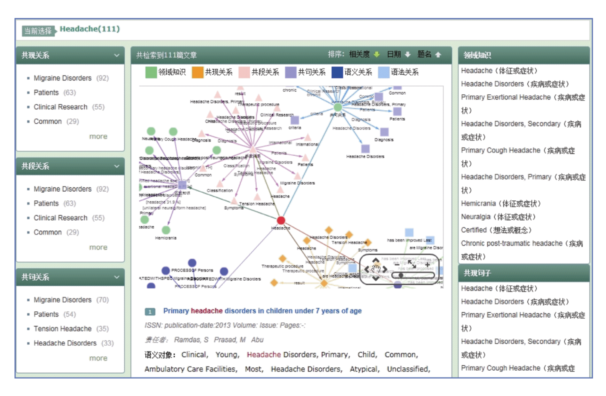


图 7　检索结果知识关系展示

这些知识及关联关系以图形化的点和边的方式展示，使用不同颜色的点代表不同类型的知识对象，不同形状表示知识之间的不同类型的关系。通过“点击–关联–展示”的导航操作方式，用户可以选择查看任意一个图上出现的知识点，并展示周边知识脉络，深入发现自己需要的知识。

示范系统能够展示检索结果中的知识关系，对科研人员判断该内容是否满足其检索需要有很大的帮助。本文认为以发现知识为先导，通过知识对象关联查看文献全文的检索方式代替传统关键词检索，对提升语义检索的精准性更有帮助。

(3) 潜在知识关系发现

潜在知识关系发现功能以检索输入的知识对象为中心，在检索结果文献中利用“共现关系索引”统计出共现、共段落和共句子的知识对象。并实现关联知识对象的分面浏览，便于科研用户从潜在的关联知识对象中发现有价值的内容，并提供导航功能筛选出这些科技文献。如图 8 上半部分所示：查询 Headache 时共现关系、共句子关系、共段落关系出现 Migraine Disorders、Clinical Research 等，对科研人员起到启示的作用。

同样，S-P-O 语义关系和句法关系分面揭示，以谓词+宾语(知识对象)的分面统计方式揭示潜在语义、句法关系。如图 8 下半部分所示：检索 Headache，通过语法关系揭示发现儿童治疗(PROCESS_OF Child)、治疗青春期(PROCESS_OF Adolescent)等深层专业领域知识的文章，通过句法关系揭示出相关治疗药物(followed Eplepsy)的研究论文等，给科研人员提供明确的知识关系启发和导向。

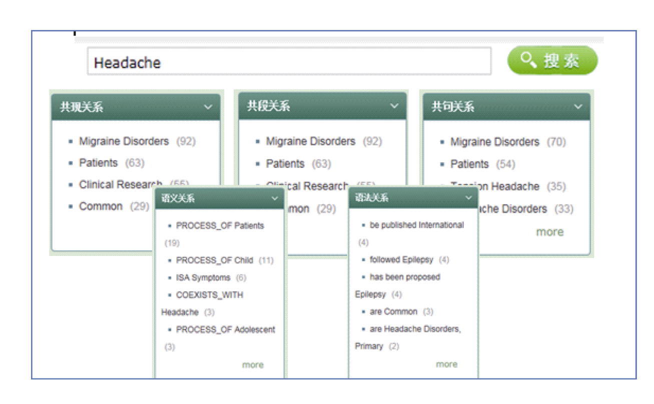


图 8　语义关系导航展示图

示范系统分面导航功能展现的知识共现关系及语义、语法关系，是根据已有文献数据统计方式揭示出来的，有助于发现文献内隐含的知识关联信息，有助于科研人员发现新的潜在的知识关系，探索学科交叉领域的新研究点，扩展科研人员的研究思路，辅助科技创新。

(4) 单篇文献语义化辅助阅读

如图 9 所示，语义化辅助阅读功能在查看单篇文献时，可以将知识对象和知识之间关系高亮展示。图 9 左侧的树形列表展示的是该篇文献中的语义知识对象，将这些语义知识对象按照类型分为不同的组，用不同颜色标示。中间主体部分是文献的文本信息，当选中某个类型的知识对象后，在中间的文本信息用该对象的颜色高亮显示出来，标示在文献中出现的位置，方便用户查阅。右侧展示该文献中计算得到的语义关系和句法关系，同样可以查看知识关系在文中的具体句子、段落的位置。

示范系统所提供的语义化辅助阅读方式，可以帮助用户直接查看知识点，直接定位知识所在的具体位置，引导读者优先阅读相关知识密集的段落和句子，从而提高对文献全文内容的阅读效率。

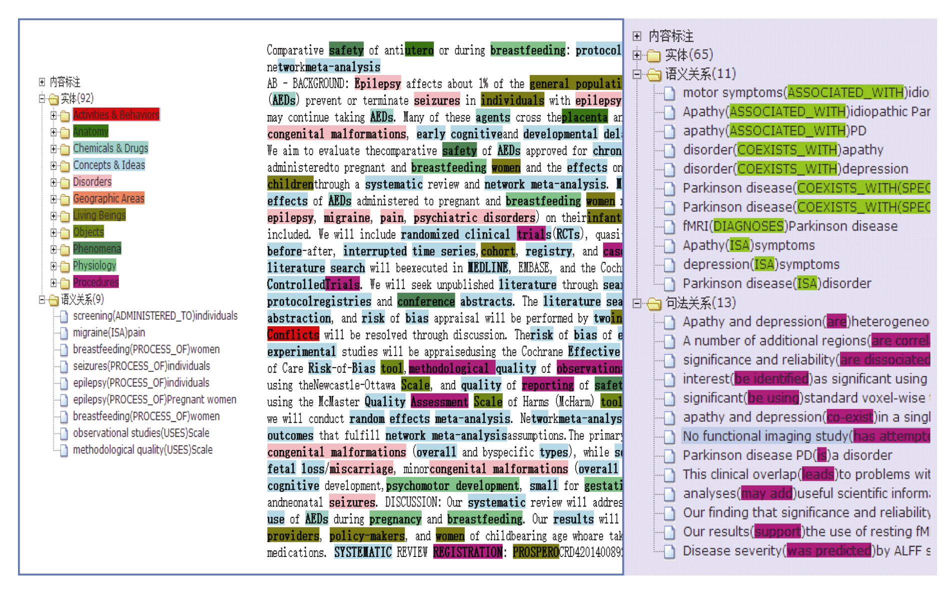


图 9　单篇文献的语义化辅助阅读展示

7　结　语

本文提出语义丰富化框架的设计模型，通过构建示范系统进一步证明语义丰富化框架的优势和可行性。从以下 4 个方面提升了文献检索效果。

(1) 语义识别技术将关键词匹配检索转变为具有语义特征的知识对象检索，提升检索精准度。辅助用户进行语义消歧，避免关键词检索出现的语义偏差。

(2) 使用相关知识对象、知识关系精准的语义表达代替列表式检索结果呈现，有助于科研人员判断检索内容是否满足其需要。

(3) 语义关联导航功能有助于发现隐含的知识关联信息，辅助科研人员发现新知识关联，探索学科交叉领域，扩展研究思路。

(4) 语义化的辅助阅读，高亮显示知识点的位置，引导读者优先阅读相关知识点密集的段落，提高文献内容的阅读效率。

在本文试验过程中，也存在一些不足之处，希望在未来的工作得以克服和改进。

(1) 句法分析得到的 S-P-O 三元组关系未能完全映射到 MeteMap 提供的 30 种规范谓词。未规范的谓词对关联导航发现功能造成一定程度的影响，后续考虑构建谓词规范词表，修改谓词语义识别算法对此改进。

(2) 知识关系揭示频繁与宽泛的上位词关联，宽泛上位词不利于帮助专业领域的科研。未来尝试通过 TF-IDF 等加权计算方法过滤频繁而宽泛的上位词，改进知识关联导航的效果。

(3) 本文试验数据集合较少，缺少大数据集上的应用测试。同时缺少对医学领域以外的应用试验以对模型进行对比评估。

参考文献：

[1] U.S. National Library of Medicine. Semantic Knowledge Representation [EB/OL]. [2016-01-13]. https://skr3.nlm.nih.gov/.

[2] Wikipedia. Knowledge Graph [EB/OL]. [2016-02-10]. https://en.wikipedia.org/wiki/Knowledge_Graph.

[3] Google Inside Search [EB/OL]. [2016-02-10]. https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html.

[4] Wolframalpha. Computational Knowledge Engine [EB/OL]. [2015-03-10]. https://www.wolframalpha.com/.

[5] Kngine. The Most Intelligent Engine [EB/OL]. [2015-03-10]. http://www.kngine.com/.

[6] SindiceTech. Enterprise Knowledge Graphs [EB/OL]. [2015-03-10]. http://www.sindicetech.com/overview.html.

[7] W3C Semantic Web. RDF [EB/OL]. [2015-06-05]. https://www.w3.org/RDF/.

[8] SindiceTech. FreeBase Distribution [EB/OL]. [2015-03-10]. http://www.sindicetech.com/freebase-distribution.html.

[9] Apache Solr [EB/OL]. [2015-06-05]. http://lucene.apache.

org/solr/.
[10] PubMed [EB/OL]. [2015-10-11]. http://www.ncbi.nlm.nih.gov/pubmed.
[11] U.S. National Library of Medicine. SemRep [EB/OL]. [2015-10-22]. https://semrep.nlm.nih.gov/.
[12] Del Corro L, Gemulla R. ClausIE: Clause-Open Information Extraction[C]//Proceedings of the the 22nd International Conference on World Wide Web. 2013:355-366.
[13] Merrill M D. Knowledge Objects[R]. USA: CBT Solutions, 1998: 1-11.
[14] U.S. National Library of Medicine. Unified Medical Language System (UMLS) [EB/OL]. [2016-01-13]. https://www.nlm.nih.gov/research/umls/ .
[15] 王颖, 张智雄, 李传席, 等. 科技知识组织体系开放引擎系统的设计与实现[J]. 现代图书情报技术,2015 (10): 95-101. (Wang Ying, Zhang Zhixiong, Li Chuanxi, et al. The Design and Implementation of Open Engine System for Scientific & Technological Knowledge Organization Systems [J]. New Technology of Library and Information Service, 2015 (10): 95-101.)
[16] UMLS. Semantic Relationships [EB/OL]. [2015-10-17]. https://www.nlm.nih.gov/research/umls/new_users/online_learning/SEM_004.html.
[17] Chakraborty A, Munshi S, Mukhopadhyay D. Searching and Establishment of S-P-O Relationships for Linked RDF Graphs: An Adaptive Approach [C]//Proceedings of International Conference on Cloud & Ubiquitous Computing & Emerging Technologies (CUBE). 2013.
[18] Matthews P H. Syntactic Relations:A Critical Survey[M]. University of CambridgePress, 2007: 3-10.
[19] U.S. National Library of Medicine. Medical Subject Headings (MeSH) [EB/OL]. [2015-06-05]. https://www.nlm.nih.gov/mesh/.
[20] U.S. National Library of Medicine. MeSH Category Tree View [EB/OL]. [2015-06-05]. https://meshb.nlm.nih.gov/#/treeSearch.
[21] MetaMap - A Tool For Recognizing UMLS Concepts in Text [EB/OL]. [2015-06-20]. https://metamap.nlm.nih.gov/.
[22] The Stanford Natural Language Processing Group. Stanford Part of Speech Tagger [EB/OL]. [2015-08-24]. https://nlp.stanford.edu/software/tagger.shtml.
[23] SPECIALIST dTagger [EB/OL]. [2015-06-20]. https://specialist.nlm.nih.gov/dTagger/.
[24] 孙坦, 刘峥. 面向外文科技文献信息的知识组织体系建设思路[J]. 图书与情报, 2013 (1): 2-7. (Sun Tan, Liu Zheng. Methodology Framework of Knowledge Organization System for Scientific & Technological Literature [J]. Library & Information, 2013(1): 2-7.)
[25] Rindflesch T C, Fiszman M. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Pprocessing: Interpreting Hypernymic Propositions in Biomedical Text [J]. Journal of Biomedical Informatics, 2003, 36(6): 462-477.

作者贡献声明:

谢靖: 语义丰富化检索模式设计及示范系统框架设计、研发, 论文主要撰写人;
王敬东: 数据处理、语义标引、关系计算工作, 论文语义计算章节撰写人;
吴振新: 论文内容结构组织, 文字编撰修订;
张智雄: 语义计算、语义索引设计思路提出;
王颖: 示范系统数据组织和图形展示方案设计;
叶志飞: 示范系统图形展示模块开发工作。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: xiej@mail.las.ac.cn。
[1] 谢靖, 王敬东. pubmed_paper.csv. PubMed提取开放获取的试验文献数据.
[2] 谢靖, 王敬东. mesh_object.csv. MeSH 提取规范后的医学知识对象数据.
[3] 谢靖, 王敬东. stkos_object.csv. STKOS 提取规范医学领域规范知识对象数据.
[4] 王敬东, 谢靖. semrep_anno.csv. SemRep 知识对象标引结果数据.
[5] 王敬东, 谢靖. semrep_spo.csv. 基于 SemRep 知识关系计算结果数据.
[6] 王敬东, 谢靖. clausIE_spo.csv. 基于 ClausIE 知识关系计算结果数据.
[7] 谢靖. solr_semantic_index.rar. 基于 Solr 语义索引数据.

收稿日期: 2017-03-03
收修改稿日期: 2017-04-08

Building Semantic Enrichment Framework for Scientific Literature Retrieval System

Xie Jing　Wang Jingdong　Wu Zhenxin　Zhang Zhixiong　Wang Ying　Ye Zhifei
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper aims to improve the scientific literature retrieval system with the help of semantic recognition and knowledge relationship computing. [**Methods**] First, we identified and extracted semantic objects from the scientific literature. Then, we calculated and established semantic relations among the objects using data-mining tools. Finally, we built semantic multidimensional index for these objects and relations, and then designed a new data organization model. [**Results**] The new system effectively identified the semantic information and improved the user experience. [**Limitations**] We need to expand the dataset used in this study and evaluate the new system in other areas. [**Conclusions**] The proposed system could retrieve more knowledge and indicate some future directions.

Keywords: Semantic Enrichment　Semantic Knowledge Organization　Semantic Relation Presentation　Multidimensional Index

Springer 和 Federica Weblearning 推出 MOOC 和教科书项目

2017 年 2 月底, Springer 与 Federica Weblearning 达成合作, Federica Weblearning 是一个致力于创新、实验和传播多媒体远程学习的学术平台。这一合作为作者和讲师在特定主题上同时编写教科书和 MOOC(Massive Open Online Course)提供了多种选择。这一举措使 Springer 能够进一步加强其在教科书出版方面的专业知识, 并为其作者和客户提供增强的教学经验。

MOOC 和教科书项目邀请了世界各地的作者从一开始就做好教科书与附带 MOOC 一起编写的计划。此外, 目前在 Federica Weblearning 平台上运行 MOOC 的讲师也有机会通过 Springer 出版相应的教科书。世界各地研究机构的作者和讲师都可以参加这一活动。

Federica Weblearning 平台上的课程广泛覆盖了大学的学科领域, 包括数学和统计学、计算机科学、工程与物理科学、生物医学和生命科学、商业与经济学、人文社会科学等。Federica Weblearning 主管 Mauro Calise 表示: “Federica Weblearning 和 Springer 在这次新的合作项目中联手展示了全球最好的国际性研究, 该项目为基于一本教科书创建 MOOC, 或基于在线课程创建一本教科书提供了独特的机会, 将科学论文的高质量与在线教育产品的交流能力相结合。”

Springer 的 MOOCs 计划执行总编辑兼项目经理 Francesca Bonadei 表示: “Federica Weblearning 平台涵盖的学科范围十分之广, 与 Springer 的广泛组合完美匹配。我们计划首先推出基于畅销作家 Bruno Siciliano 撰写的图书而制作的新 MOOC, 同时, 我们期待与 Federica Weblearning 的讲师一起帮助他们出版课程所附带的教科书。”

(编译自: https://www.springer.com/gp/about-springer/media/press-releases/corporate/moocs-and-books-initiative-launched-by-springer-and-federica-weblearning/12241436)

(本刊讯)